Security of Infocommunication Systems and Internet of Things

2024 Vol 2, No 1

https://doi.org/10.31861/sisiot2024.1.01002

Received 20 May 2024; revised 22 August 2024; accepted 28 August 2024; published 30 August 2024

Advanced Data Aggregation in Online Education: a Contextual Web Parser Approach

Kostiantyn Foksha[1] and Ganna Zavolodko[2,*]

[1]Department of Multimedia and Internet technologies and systems, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
[2]Department of Information system, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

*Corresponding author (E-mail: anna.zavolodko@khpi.edu.ua)

ABSTRACT The paper presents a web aggregator system for collecting, filtering, and classifying data from educational platforms, focusing on online courses. It describes the development and testing of a system that uses contextual search to help users find courses matching their interests and knowledge level, while also handling spelling errors. The system's effectiveness is established through tests demonstrating its capability for rapid data collection and update, providing accurate and relevant results. The paper details the system's three-tier structure: data aggregation, user filtering, and user-system interaction for tailored course recommendations. The development involves a Python web server, a MariaDB database, a parser for non-formal education platforms, and a web application for client data presentation. In this paper also highlight the system's scalability and potential for integration with other educational platforms. Emphasize the importance of continuous updates to the database for maintaining relevance in a rapidly evolving online education landscape. Additionally, the paper discusses future enhancements, including the implementation of advanced machine learning algorithms for improved search accuracy and personalization, emphasizing the system's ongoing evolution to meet the dynamic needs of online learners.

KEYWORDS web aggregator, online education platforms, course recommendation systems, contextual search, data collection and filtering.

I. INTRODUCTION

In the ever-evolving landscape of education, the proliferation of online learning platforms has generated an unprecedented volume of educational content. This requires advanced systems for efficient data aggregation and management. This study focuses on the development of a sophisticated web aggregator that uses cutting-edge technologies, such as a Python-based web server and a MariaDB database, to systematically collect, filter and present educational data. The research highlights the importance of non-formal education platforms and is at the forefront of educational technology. It offers a transformative tool that promises to improve user engagement and learning outcomes. It represents a step towards personalised and accessible learning experiences, given the system's ability to adapt to the dynamic nature of online education, and its potential for scalability and integration with machine learning algorithms. By providing actionable insights and fostering an enriched educational ecosystem, the anticipated outcomes of this research could make a significant contribution to the field. The development uses a Python web server, a MariaDB database that stores all information about users and courses, a parser for collecting data from non-formal education platforms that then updates the server with new aggregated data, and a web application for providing data to the client. It has been developed by analysing and synthesising sources [1-12].

The analysis in this article was conducted using ChatGPT as a tool that allows to simplify routine work. The use of ChatGPT in the study highlights its usefulness in processing and analysing information from various online educational platforms [13]. This integration demonstrates the potential of artificial intelligence tools such as ChatGPT to improve research methodologies and provide in-depth insights into subject areas, especially in the emerging field of online education and web data aggregation.

Algorithms for aggregating data from the Internet have not yet been studied, nor have possible approaches to this been analyzed. The analysis of algorithms in this work is based on various types of sites as well as their methods of storing and providing information in context of education platforms. In order for users to access courses from other learning platforms on platformoEDU, courses must be stored in the platform database. In order to receive courses from different platforms in platformoEDU, you need to use an information aggregator that will scan data from other platforms and update the relevant information. The general scheme of the aggregator is shown in Figures 1, 2).

Among the requirements for the program are the following main requirements:

1. Data collection validation – the program should check the validity of the data.
2. Characterization – course data such as topic, difficulty level, duration, etc.
3. Error handling – the program should be able to handle errors, which should be recorded in the error knowledge base to check the problem later.
4. Speed and efficiency – the selected tools for

implementing the program should have the highest efficiency among analogues.

5. Availability of data on the work process – the program should provide the ability to display the processes that occur in it during execution.

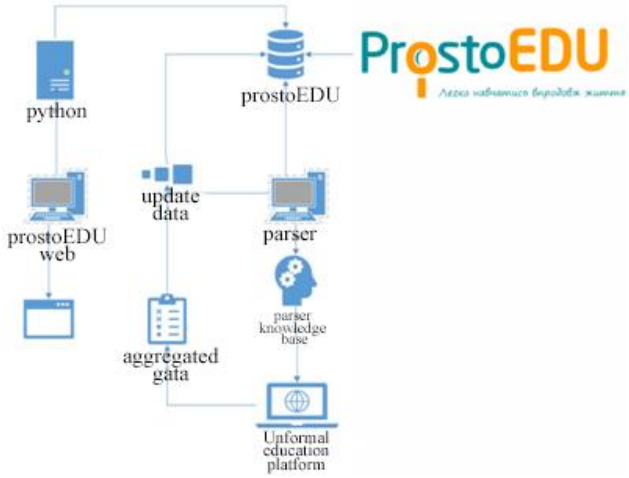


FIG. 1. Scheme of the web analyzer work.

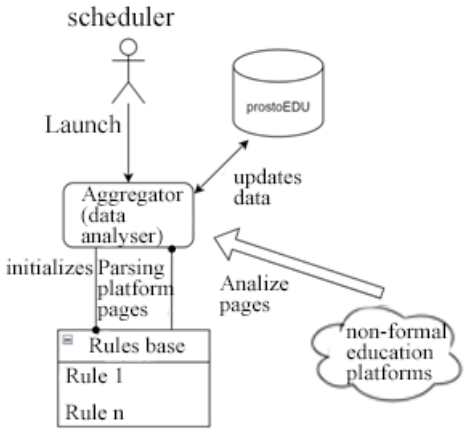


FIG. 2. Scheme of the data aggregator.

II. ALGORITHM OF THE WEB AGGREGATOR

In the dynamic and ever-expanding realm of web data extraction, the development and implementation of effective web parsers are crucial. These tools are designed to systematically navigate, extract, and process data from various websites, each presenting unique structural and content challenges. This section delves into the intricacies of web parsers, outlining different strategies tailored to the specific architectures of websites. We explore several schemes: a general parser workflow, a parser operating from the main page, another working with the /sitemap of a website, and lastly, a parser that navigates through categories.

Each of these approaches is meticulously crafted to address the distinct characteristics and layouts of websites, ensuring a comprehensive and efficient data extraction process. From simple, main-page-focused structures to more complex, category-based or /sitemap-oriented designs, these parsing strategies demonstrate the adaptability and precision required in modern web data collection. Understanding these schemes provides invaluable insights into the mechanics of web scraping, a process integral to the vast domain of data-driven analysis and decision-making in today's digital world.Depending on the structure of the website, the approach to collecting data from the platform also changes - some sites store links in /sitemap, some provide a list of pages in xml, others in links to categories, etc.

Data processing algorithms depend on the content of the page, the structure of the web page, the type and format of data, the volume and complexity of data, data availability, website limitations, and other factors, but we can distinguish the general scheme of a web parser.

Fig. 3 illustrates the general workflow of a web parser. The process begins with obtaining the website's URL, followed by the acquisition of a list of processed pages. This is key to ensure that the parser does not revisit pages already present in the database. Should the list reveal unprocessed pages, the system proceeds to collect information from these pages. Upon completing the collection from all pages, the database is updated accordingly. The process concludes once the update is finished, signifying the end of the cycle. This workflow adapts to various website structures and data formats, ensuring flexibility in data collection and processing.

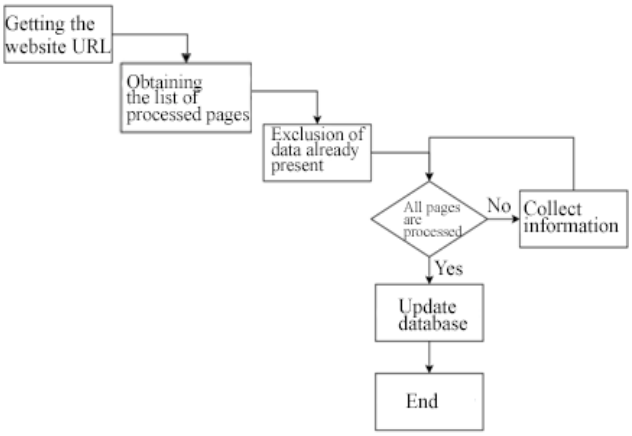


FIG. 3. General scheme of the web parser workflow.

A. Scheme of the web parser from the main page. This scheme is suitable for collecting data from a website where the links to the data to be collected are located on the main page and are not available at the /sitemap URL. The parser collects links to pages and then collects data from these pages. The parser scheme is shown in Fig. 3. The code of the program that implements the parser according to this scheme is shown in Fig. 4

Fig. 4 presents a scheme for a web parser designed to operate from the main page of a website. This specific scheme is tailored for instances where the data links are located directly on the main page rather than through a /sitemap URL. The parser starts by getting the URL, then checks if all the pages have been collected. If not, it collects data from the current page. Once all pages are collected, the parser extracts the data, updating it accordingly. After the update, the process ends. This scheme is particularly efficient for websites with a simple structure where all necessary links are immediately accessible from the homepage.

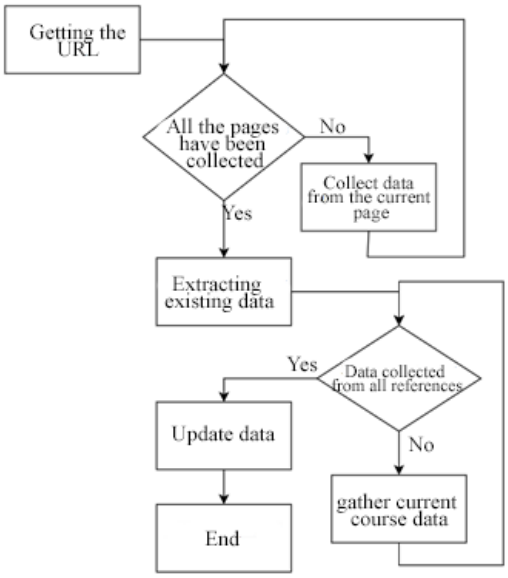


FIG. 4. Parser scheme from the main page.

B. Scheme of how a web parser works with /sitemap. This scheme is suitable for a site where the data to be collected is located on the links located in the /sitemap branch of the site. The parser scheme is shown in Fig. 5.

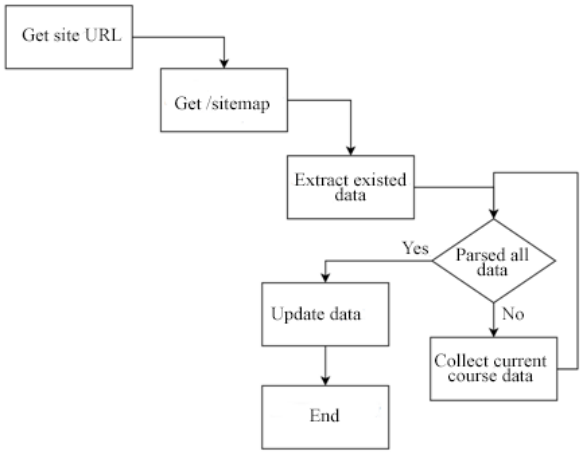


FIG. 5. Scheme of the parser with /sitemap.

Fig. 5 depicts a workflow for a web parser that navigates a site's /sitemap. This approach is optimized for websites that organize the data to be collected within the /sitemap directory. The process begins with retrieving the site's URL, followed by accessing the /sitemap. It then extracts existing data to determine what needs to be updated. The parser checks if all necessary data has been parsed. If not, it proceeds to collect the current course data. Once all data is collected and parsed, it updates the database, concluding the operation with the end of the data update. This methodical approach ensures a thorough and systematic collection of data.

C. Scheme of web parser operation by category. This scheme is suitable for collecting data from a website where the data is linked to categofies. The parser scheme is shown in Fig. 6.

Fig. 6 outlines a parser workflow designed for websites where data is organized by category. The process begins

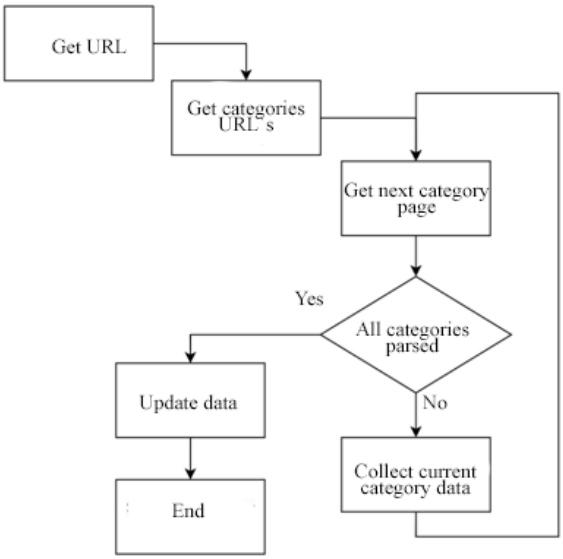


FIG. 6. Parser scheme by category.

with obtaining the main URL, then it retrieves the URLs for each category. The parser systematically visits each category, checking if there are more pages within the category to parse. If not, it collects data from the current category page. This cycle repeats until data from all categories has been collected. Once the collection is complete, the data is updated in the system, and the process ends. This scheme is particularly useful for structured websites with distinct categories of information.

III. DATA OUTPUT

In this crucial section, we delve into the data output phase of the web aggregator's operation, a fundamental aspect of web scraping and data collection processes. This phase not only represents the culmination of the data extraction effort but also provides critical insights into the efficiency and effectiveness of the web aggregator. We will explore various facets of this process, starting with the data collection process, where we analyze the output displayed during the aggregator's operation. This includes details such as the source of data, the volume of data already existing in the knowledge base, the amount of new data collected, and the time taken for each operation.

Illustrative Figures, such as Fig. 7 and Fig. 8, will shed light on the initial and subsequent executions of the program, revealing how the aggregator adapts and responds to the changing data landscape across different platforms. We will discuss the implications of these results, including the efficiency of data collection and the update mechanisms in place.

Following this, we will delve into the structuring and management of the knowledge base. This includes the transformation of raw data into a structured and easily navigable format, as exemplified in Fig. 9, which showcases the data organized in an Excel spreadsheet. This step is crucial for ensuring that the data is not only collected but also presented in a manner that is accessible and useful for further analysis or application.

Finally, we will examine how the aggregated data is integrated into the platform, with a focus on the user

interface and data accessibility, as demonstrated in Fig. 10. This section will highlight the practical applications of the aggregated data and the ease with which users can interact with and benefit from the collected information.

Overall, this section aims to provide a comprehensive understanding of the data output stage in web aggregation, emphasizing the importance of efficient data handling, organization, and presentation in making the collected data valuable and actionable.

A. Data collection process. When the aggregator is tunning, the console displays information that shows the result of the program. It displays the site from which the data is collected, the amount of data from this platform that is already in the knowledge base, the amount ofdata that was collected, and the time it took the program to complete the work. The result of the first lun of the program with a limited amount of data is shown in Fig. 7.

```
PS D:\study\WebScrapper> & "C:/Program Files/python311/python.exe" d:study/WebScrapper/main.py
Parsing https://coursera.org
DB has 0 courses from coursera
Done. Coursera parsed with 1764 courses. Total of 1764 courses in DataBase. Parsing time: 422sec

Parsing https://sololearn.com
DB has 0 courses from sololearn
Done. Sololearn parsed with 27 courses. Total of 27 courses in DataBase. Parsing time: 36sec

Parsing https://alison.com
DB has 0 courses from alison
Done. Alison parsed with 30 courses. Total of 30 courses in DataBase. Parsing time: 115sec

Parsing https://edx.com
DB has 0 courses from edx
Done. Edx parsed with 64 courses. Total of 64 courses in DataBase. Parsing time: 506sec
```

FIG. 7. The result of running the program for the first time.

Upon the initial execution of the web aggregator program, it was observed that the database started empty for each platform. The program successfully collected data from Coursera, Sololearn, Alison, and edX, adding the new course information to the knowledge base. The duration of the program's operation varied, reflecting the different data collection methods adapted for each platform's unique structure. A subsequent run of the program, with limitations on the volume of data still in place, would yield further insights into the aggregator's efficiency and the potential incremental additions to the database. The outcomes of this second run are presented in Fig. 8, which is not included here.

```
PS D:\study\WebScrapper> & "C:/Program Files/python311/python.exe" d:study/WebScrapper/main.py
Parsing https://coursera.org
DB has 1764 courses from coursera
Done. Coursera parsed with 0 courses. Total of 1764 courses in DataBase. Parsing time: 68sec

Parsing https://sololearn.com
DB has 27 courses from sololearn
Done. Sololearn parsed with 0 courses. Total of 27 courses in DataBase. Parsing time: 10sec

Parsing https://alison.com
DB has 30 courses from alison
Done. Alison parsed with 30 courses. Total of 60 courses in DataBase. Parsing time: 108sec

Parsing https://edx.com
DB has 64 courses from edx
Done. Edx parsed with 0 courses. Total of 64 courses in DataBase. Parsing time: 31sec
```

FIG. 8. The result of the second execution of the program.

Fig. 8 displays the console output after the second execution of the web aggregator program. It shows the parsing process of the same educational platforms as in the first run. However, this time, no new courses were found on Coursera and edX, indicating that the previous data from these platforms was up-to-date. In contrast, additional courses were found for Alison, doubling the total count in the database for this platform. The parsing times for each site are also noted, with Coursera taking the longest at 68 seconds and Sololearn the shortest at 10 seconds, showcasing the system's capability to efficiently check for new data and update the database accordingly.

B. Knowledge base. Comparing the execution time with the first lun, it is noted that some programs were faster but did not collect any data - this is because the data selection algorithm collected data according to the test mode constraints and then removed from the list of data those that are already in the knowledge base, which demonstrates that the programs select only data for processing that does not yet have infonnation, and that all data was collected successfully on the first attempt. The third program managed to collect the same amount of data as the first time, because the collection algorithm includes collecting all links from sitemap, removing existing links, and collecting data about new ones.

C. Update data in the knowledge base. After collecting the data, we will output the data to an excel file for easy viewing. The data obtained are shown in Fig. 9.

Fig. 9 exhibits a snapshot of the knowledge base output to an Excel file after data collection. The

	title	description	author	link	image_link	durations	rate	rate_count	students	document	language	price	difficulty	platform_list	form_nat	tags
0	Japanese This free c	NPTEL		https://all	https://cd	1.5-3 Hour	-1	94	30219	Alison Cer	N/A	Free	For all	https://all	alison	['Japanese Language', 'Travel', 'Vocabulary']
1	Introducti This free c	Saylor Acc		https://all	https://cd	No duratio	-1	12	17057	Alison Cer	N/A	Free	For all	https://all	alison	['Digital Marketing', 'Marketing', 'Sales', 'Advertising', 'Marketing Strategy']
2	Diploma i	With this	Blawiz Ltd	https://all	https://cd	6-8 Hours	-1	8	8752	Alison Cer	N/A	Free	For all	https://all	alison	['Manufacturing', 'Product Management', 'Quality Assurance', 'Operations', 'Branding', 'Food and Beverage', 'Innovation']
3	Protect Yo	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	95	11623	Alison Cer	N/A	Free	For all	https://all	alison	['Cyber Security', 'Information Security', 'Data Security']
4	Introducti	This cours	Advance L	https://all	https://cd	1.5-3 Hour	-1	430	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Photography', 'Digital Photography', 'DSLR', 'Camera']
5	English Gr	This Englis	Advance L	https://all	https://cd	1.5-3 Hour	-1	118	28845	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Vocabulary', 'English Language', 'Grammar', 'English Speaking', 'Vocabulary', 'Intermediate English']
6	Introducti	This free c	Microsoft	https://all	https://cd	1.5-3 Hour	-1	9	8497	Alison Cer	N/A	Free	For all	https://all	alison	['Web Design', 'JavaScript', 'Visual Studio', 'HTML', 'Bootstrap', 'UI']
7	Introducti	This leade	Advance L	https://all	https://cd	1.5-3 Hour	-1	214	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Supervision', 'Motivation', 'Communication Skills', 'Transition Year', 'Transition Year Enterprise', 'Homeschool']
8	Wind Enel	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	67	17282	Alison Cer	N/A	Free	For all	https://all	alison	['Renewable Energy', 'Environmental Management', 'Climate Change', 'SDG 7 Affordable And Clean Energy', 'SDG 13 Climate Action']
9	Beginner i	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	113	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Photography', 'Photoshop', 'Adobe', 'Photo Editing', 'Digital Photography', 'Lightroom', 'DSLR', 'Camera', 'Transition Year', 'Transition Year Creativity']
10	Human Re	This free c	NPTEL	https://all	https://cd	1.5-3 Hour	-1	116	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Labour Law', 'Productivity', 'Conflict Management', 'Employment Law', 'Human Resource Management']
11	Introducti	This free c	Microsoft	https://all	https://cd	1.5-3 Hour	-1	6	10853	Alison Cer	N/A	Free	For all	https://all	alison	['Programming', 'Web Development', 'JavaScript', 'AngularJS', 'TypeScript', 'UI', 'MVC', 'Server']
12	Supervisio	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	112	27909	Alison Cer	N/A	Free	For all	https://all	alison	['Supervision', 'Communication Skills']
13	Beginner i	This free c	Advance L	https://all	https://cd	3-4 Hours	-1	534	32767	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Speaking']
14	Financial i	This free c	Entreprer	https://all	https://cd	1.5-3 Hour	-1	202	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Accounting', 'Bookkeeping', 'Financial Accounting', 'Basic Accounting', 'Auditing', 'Cash Flow', 'SDG 1 No Poverty']
15	3D Printin	This free c	Channel 9	https://all	https://cd	1.5-3 Hour	-1	15	11442	Alison Cer	N/A	Free	For all	https://all	alison	['Manufacturing', '3D Printing', 'Windows 10', 'Transition Year Top Ups and Tasters']
16	Intermedi	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	36	26769	Alison Cer	N/A	Free	For all	https://all	alison	['Photography', 'Photo Editing', 'Digital Photography']
17	Beginner i	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	242	32767	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Vocabulary', 'English Language', 'Grammar', 'English Speaking', 'Basic English', 'Vocabulary']
18	Building a	These day	Microsoft	https://all	https://cd	1.5-3 Hour	-1	1	9643	Alison Cer	N/A	Free	For all	https://all	alison	['Web Development', 'Web Design', 'JavaScript', 'Visual Studio', 'HTML', 'Bootstrap', 'User Interface']
19	Supervisio	This free c	Advance L	https://all	https://cd	4-5 Hours	-1	27	18958	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Supervision', 'Motivation', 'Communication Skills', 'Conflict Management', 'Goal Setting', 'Bullying', 'Performance Management']
20	Internatio	This free c	NPTEL	https://all	https://cd	1.5-3 Hour	-1	35	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Supervision', 'Ethics', 'Employment Law', 'Human Resource Management', 'Employee Engagement']
21	Solar Ener	This cours	Advance L	https://all	https://cd	1.5-3 Hour	-1	161	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Geography', 'Renewable Energy', 'Solar Energy', 'Climate Change', 'SDG 7 Affordable And Clean Energy', 'SDG 13 Climate Action']
22	Introducti	This free c	Microsoft	https://all	https://cd	1.5-3 Hour	-1	23	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Programming', 'Software Development', 'Computer Science', 'Coding', 'C++']
23	Supervisio	This free c	Advance L	https://all	https://cd	3-4 Hours	-1	73	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Supervision', 'Communication Skills', 'Conflict Management', 'Small Business', 'Bullying']
24	Beginner i	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	205	32767	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Vocabulary', 'English Language', 'Travel', 'Grammar', 'English Speaking', 'Basic English', 'Vocabulary']
25	Advanced	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	61	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Photography', 'Photoshop', 'Photo Editing', 'Digital Photography', 'DSLR', 'Camera', 'Transition Year', 'Transition Year Creativity']
26	Cash Flow	This free c	Entreprer	https://all	https://cd	1.5-3 Hour	-1	29	17841	Alison Cer	N/A	Free	For all	https://all	alison	['Accounting', 'Business Analysis', 'Auditing', 'Cash Flow', 'Transition Year Enterprise']
27	Diploma i	This diplo	NPTEL	https://all	https://cj	10-15 Hou	-1	98	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Japanese Language', 'Transition Year Top Ups and Tasters', 'Languages']
28	Introducti	On your w	Advance L	https://all	https://cd	4-5 Hours	-1	28	8451	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Business Analysis', 'Strategic Planning', 'Innovation', 'Strategic Thinking', 'Leadership Skills']
29	C++ Progr	This free c	Microsoft	https://all	https://cd	1.5-3 Hour	-1	4	15479	Alison Cer	N/A	Free	For all	https://all	alison	['Programming', 'Computer Science', 'Coding', 'C++', 'Programming Languages']
30	Diploma i	This free c	Advance L	https://all	https://cj	10-15 Hou	-1	134	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Supervision', 'Communication Skills', 'Mentoring', 'Bullying', 'Performance Management', 'Leadership Skills']
31	Diploma i	This cours	Openstax	https://all	https://cj	6-10 Hour	-1	1506	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Customer Service', 'Hospitality Management', 'Hospitality', 'Retail', 'Food Service', 'Retail Apprenticeships']
32	Fundamen	This free c	XSIQ	https://all	https://cj	1.5-3 Hour	-1	871	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Human Resources', 'Employment Law', 'Human Resource Management']
33	Diploma i	This free c	XSIQ	https://all	https://cd	10-15 Hou	-1	824	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Health Management', 'Health and Fitness', 'Nutrition', 'Physical Therapy', 'Health', 'Healthcare', 'HIV', 'Physiotherapy', 'Diet', 'Disability', 'SDG 3 Good Health And Wellbeing', 'SDG
34	Diploma i	This diplo	Global Tei	https://all	https://cj	10-15 Hou	-1	837	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Teaching', 'Classroom Management', 'Educational Psychology', 'SDG 4 Quality Education']
35	Beginner i	This cours	Advance L	https://all	https://cd	1.5-3 Hour	-1	88	16019	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Vocabulary', 'English Language', 'Grammar', 'English Speaking', 'Basic English', 'Vocabulary']
36	Diploma i	This cours	MEDCoE	https://all	https://cj	10-15 Hou	-1	1024	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Anatomy', 'Science', 'Biology', 'Physiology', 'Human Anatomy']
37	Diploma i	This free c	XSIQ	https://all	https://cj	10-15 Hou	-1	1018	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Business Law', 'Contract Law']
38	Diploma i	This free c	XSIQ	https://all	https://cj	6-10 Hour	-1	2841	32767	Alison Cer	N/A	Free	For all	https://all	alison	['English Vocabulary', 'English Speaking', 'Vocabulary']
39	Beginner i	This free c	Advance L	https://all	https://cd	1.5-3 Hour	-1	16	5847	Alison Cer	N/A	Free	For all	https://all	alison	['Customer Service', 'Digital Marketing', 'Retail', 'Transition Year Enterprise', 'Retail Apprenticeships']
40	Diploma i	This free c	Global Tei	https://all	https://cj	6-10 Hour	-1	925	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Programming', 'Information Systems', 'Databases', 'Security', 'Database Management', 'Decision Making', 'IT Management', 'Information Management', 'Programm
41	Fundamen	This free c	XSIQ	https://all	https://cd	1.5-3 Hour	-1	641	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Project Management']
42	Ruby on R	This free c	Microsoft	https://all	https://cd	3-4 Hours	-1	4	6284	Alison Cer	N/A	Free	For all	https://all	alison	['App Development', 'Programming', 'Web Development', 'Web Design', 'Ruby', 'Ruby on Rails', 'MVC']
43	Diploma i	This free c	Openstax	https://all	https://cj	15-20 Hou	-1	703	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Food Safety', 'Cooking', 'Restaurant Management', 'Hygiene', 'Food Service', 'SDG 2 Zero Hunger']
44	Diploma i	This free c	XSIQ	https://all	https://cj	6-10 Hour	-1	3124	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Science', 'Psychology', 'Psychotherapy', 'Clinical Psychology']
45	Music The	This free c	Open Yale	https://all	https://cd	1.5-3 Hour	-1	19	20044	Alison Cer	N/A	Free	For all	https://all	alison	['Music', 'Music Theory', 'Transition Year Creativity']
46	Introducti	This free c	Global Tei	https://all	https://cj	1.5-3 Hour	-1	801	32767	Alison Cer	N/A	Free	For all	https://all	alison	[]
47	Basics of	This free c	XSIQ	https://all	https://cd	1.5-3 Hour	-1	316	11675	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Change Management', 'Business Management', 'Corporate Finance', 'Corporate Governance']
48	Diploma i	This free c	Advance L	https://all	https://cd	10-15 Hou	-1	1708	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Programming', 'Operating Systems', 'C Programming', 'Computer Science', 'Programming Languages']
49	Financial i	This free c	Christine '	https://all	https://cd	1.5-3 Hour	-1	26	23225	Alison Cer	N/A	Free	For all	https://all	alison	['Investment', 'Finance', 'Personal Finance', 'Financial Management', 'Insurance', 'Mortgage', 'Cash Flow', 'Debt', 'Transition Year Enterprise']
50	Diploma i	This free c	MEDCoE	https://all	https://cj	10-13 Hou	-1	1988	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Nutrition', 'Nursing', 'SDG 8 Decent Work And Economic Growth']
51	Diploma i	Learn abo	Openstax	https://all	https://cj	10-15 Hou	-1	487	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Hospitality Management', 'Cooking', 'Hotel Management', 'Restaurant Management', 'Food and Beverage', 'Hospitality', 'Catering']
52	Diploma i	With this	Advance L	https://all	https://cd	6-10 Hour	-1	1807	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Health Management', 'Health and Safety', 'Compliance', 'Ergonomics']
53	Fundamen	This free c	XSIQ	https://all	https://cd	1.5-3 Hour	-1	859	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Accounting', 'Bookkeeping', 'Financial Analysis', 'Auditing', 'SDG 1 No Poverty']
54	Beginner i	This basic	Advance L	https://all	https://cd	1.5-3 Hour	-1	83	26674	Alison Cer	N/A	Free	For all	https://all	alison	['English Grammar', 'English Vocabulary', 'English Language', 'Grammar', 'English Speaking', 'Basic English', 'Vocabulary']
55	Diploma i	This diplo	XSIQ	https://all	https://cj	10-15 Hou	-1	4960	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Management', 'Project Management', 'Time Management']
56	Diploma i	This free c	Brian Lind	https://all	https://cj	10-15 Hou	-1	724	32767	Alison Cer	N/A	Free	For all	https://all	alison	['Health and Fitness', 'Nutrition', 'Health', 'Wellness', 'Human Nutrition', 'SDG 2 Zero Hunger']

FIG. 9. The data obtained.

spreadsheet includes detailed columns such as title, description, author, link, image, duration, rate, student count, document language, price, difficulty level, platform, and tags. This format allows for an organized and accessible presentation of the collected data, showcasing the diversity of courses extracted from the educational platforms. The data spans various subjects and shows key information, facilitating easy sorting and analysis.

The data obtained as a result of the relationship is then entered into the platform by the operator. For convenience, let's add the name of the platform to the course name to see data from all platforms. The result of adding courses to the platform is shown in Fig. 10.

Fig. 10 displays the database interface where course data, collected from various online educational platforms, has been entered. Each course entry includes the platform's name prefixed to the course title, enabling a clear and organized view of the aggregated data across multiple sources. This naming convention ensures that users can easily identify the origin of each course and facilitates efficient navigation through the database.

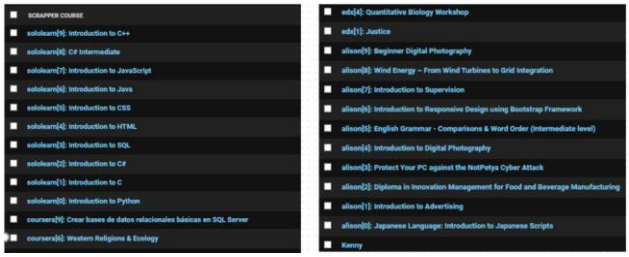


FIG. 10. Database.

IV. TESTING

To compare the colTectness of the program, 3 tests were perfonned – from a small amount of data (the limitation depends on the data processing method), to analysis over the entire intelval. The table comparing the program results is shown in Table 1. A comparison ofthe program speed per data unit per second is shown in Fig. 11.

Table 1 provides a detailed comparison of the results from three different tests conducted to evaluate the performance of the developed aggregator across various platforms including Coursera, Alison, Sololearn, and edX. For each attempt, the table lists the time taken in seconds, the amount of data collected (measured in units), the time efficiency calculated as time per unit, and the number of errors encountered. Over the three tests, we see a variation in the time efficiency and accuracy, with the first attempt being the most error-free across all platforms. As the volume of data increases in subsequent attempts, there's a noticeable increase in both time taken and errors, suggesting a correlation between the dataset size and the likelihood of errors. This table serves as a benchmark for assessing the aggregator's efficiency and reliability.

The bar graph in Fig. 11 compares the operational speed of the web aggregator across four different educational platforms: Coursera, edX, Sololearn, and Alison. Each platform is represented by a different color, and the height of each bar corresponds to the time

TABLE 1. Comparsion of the results of the developed aggegator.

Attempt	Characteristics	Platform			
		Coursera	Alison	Sololearn	Edx
1	Time (s)	17	35	8	204
	Amount of data (units)	60	10	5	32
	Time per unit (s)	0.28	3.5	1.6	6.37
	Errors (units)	0	0	0	0
2	Time (s)	422	115	14	506
	Amount of data (units)	1764	30	15	64
	Time per unit (s)	0.24	3.83	0.9	7.9
	Errors (units)	2	0	0	0
3	Time (s)	4655	13647	12	3056
	Amount of data (units)	8766	3791	27	456
	Time per unit (s)	0.53	3.59	0.44	6.7
	Errors (units)	15	10	0	1

efficiency in processing the data. From the graph, we can observe that Coursera and Sololearn exhibit a higher speed of operation compared to Alison and edX, indicating a more efficient data processing capability in the former two. The comparison highlights the variance in processing times and could be used to identify performance benchmarks for each platform.

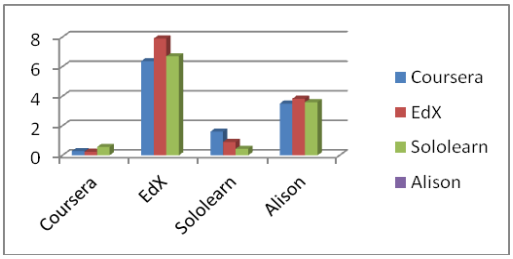


FIG. 11. Compalison of speed of operation.

Let's test adding 1000 courses to the database. The result is shown in Fig. 12.

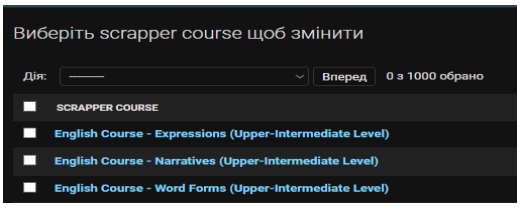


FIG. 12. The result of adding courses to the database.

As a result of adding data to sites, all data is added successfully without exception

V. CONCLUSIONS

The result entitled has shed light on the effectiveness of the developed Web aggregator for collecting, filtering, and categorizing information from varied online learning environments. It emphasizes the successful execution of distinct parser configurations that have been fine-tuned to accommodate diverse web architectures, leading to

proficient gathering and updating of data. The aggregator's capability to neatly compile this information within the knowledge base and via the user interface is also showcased. The paper underscores the necessity for constant updates and the possible incorporation with machine learning technologies to boost customization and precision, thereby demonstrating the system's flexibility and growth potential within the dynamic realm of online education. Test findings underline the aggregator's speed and dependability, affirming its value in supporting non-formal education platforms. In conclusion, the research presents several strategies on how to synthesize and manage data from websites with varying storage and presentation methods.

AUTHOR CONTRIBUTIONS

K.F., G.Z. – conceptualization, methodology, investigation, writing (original draft preparation), writing (review and editing).

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] Introduction to the DOM. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model/Introduction.

[2] Robie J. What is the Document Object Model? [Online]. Available: https://www.w3.org/TR/WD-DOM/introduction.html.

[3] Matyash D. Navishcho potribni sayty-ahrehatory, chomu Google yikh tak lyubyt'? [Online]. Available: https://jam.in.ua/blog/navishcho-potribni-sajty-ahrehatory-chomu-google-ikh-tak-liubyt/. [in Ukrainian]

[4] Shcho take kraulinh i yak keruvaty robotamy. [Online]. Available: https://www.bizmaster.xyz/2019/04/scho-take-krauling-i-yak-keruvaty-robotamy.html. [in Ukrainian]

[5] Henderson A. "15 Best FREE Website Crawler Tools & Software (2023 Update)." [Online]. Available: https://www.guru99.com/web-crawling-tools.html.

[6] Digital Commerce Intelligence. [Online]. Available: https://www.dexi.io/.

[7] Horobtsov V. Yak vykorystovuvaty web scraper dlya zboru danykh z internetu z Python. [Online]. Available: https://dou.ua/forums/topic/43070/. [in Ukrainian]

[8] What is an API? [Online]. Available: https://www.ibm.com/topics/api.

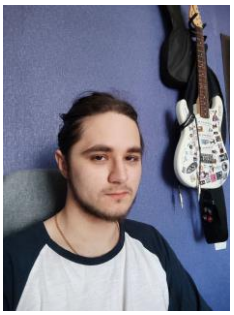
[9] Whitehead C. T. What Is an RSS Feed? (And Where to Get It). [Online]. Available: https://www.lifewire.com/what-is-an-rss-feed-4684568.

[10] Requests: HTTP for Humans™. [Online]. Available: https://docs.python-requests.org/en/latest/index.html.

[11] Daityari S. "App & Browser Testing Made Easy." [Online]. Available: https://www.browserstack.com/guide/python-selenium-to-run-web-automation-test.

[12] Web Scraping with Selenium and Python Tutorial + Example Project. [Online]. Available: https://scrapfly.io/blog/web-scraping-with-selenium-and-python/.

[13] F.M.M. Morrison, N. Rezaei, A.G. Arero, V. Graklanov, S. Iritsyan, M. Ivanovska, R. Makuku, L.P. Marquez, K. Minakova, L.P. Mmema, P. Rzymski, G. Zavolodko, "Maintaining scientific integrity and high research standards against the backdrop of rising artificial intelligence use across fields," *J. Med. Artif. Intell.*, vol. 6, 2023.



Kostiantyn Foksha

NTU "KhPI" student, computer scientist, IEEE Extreme 17 participant.

ORCID ID: 0000-0001-7119-0401

Ganna Zavolodko

PhD in Informstion technology, Associate Professor of NTU "KPI", IEEE Senior.

ORCID ID: 0000-0003-0000-8910

Просунута агрегація даних в онлайн-освіті: підхід контекстного веб-парсеру

Костянтин Фокша[1], Ганна Заволодко[2,*]

[1]МІТС, КН, НТУ «ХПІ», Харків, Україна
[2]СІ, ІКМ, НТУ «ХПІ», Харків, Україна

*Автор-кореспондент (Електронна адреса: E-mail: anna.zavolodko@khpi.edu.ua)

АНОТАЦІЯ У статті представлено структуру веб-агрегатора для збору, фільтрації та класифікації даних з освітніх платформ, зосереджених на онлайн-курсах. Показано архітектуру та результати тестування розробки, яка агрегує данні для системи, яка використовує контекстний пошук, щоб допомогти користувачам знайти курси, які відповідають їхнім інтересам та рівню знань, а також обробляє орфографічні помилки. Описано основні архітектурні елементи розробленого модулю. Ефективність системи підтверджується тестами, які демонструють її здатність до швидкого збору та оновлення даних, надання точних і релевантних результатів. У статті детально описано трирівневу

структуру системи: агрегація даних, фільтрація користувачів та взаємодія користувача з системою для надання індивідуальних рекомендацій щодо курсів. Розробка включає веб-сервер на мові Python, базу даних MariaDB для зберігання результатів парсингу, парсер, який оснований на використанні бібліотек для платформ неформальної освіти та модуль міграції для кросплатформеного веб-додатку для представлення даних клієнтам. У цій статті також підкреслимо масштабованість та потенціал продуктового рішення для інтеграції з іншими освітніми платформами. Підкреслюється важливість постійного оновлення бази даних для підтримання її актуальності у швидкозмінному ландшафті онлайн-освіти. Для цього пропонується зробити модулі для авто адаптації під змінні умови. Крім того, в документі обговорюються майбутні вдосконалення, включаючи впровадження передових алгоритмів машинного навчання для підвищення точності пошуку та персоналізації, підкреслюючи постійну еволюцію системи для задоволення динамічних потреб онлайн-учнів. Таким чином, стаття підсумовує досвід розробки рішення для ефективної взаємодії з освітніми ресурсами, спрямованого на забезпечення якісного підбору навчальних курсів і підвищення зручності користування онлайн-освітніми платформами.

КЛЮЧОВІ СЛОВА веб-агрегатор, платформи онлайн-освіти, контекстний пошук, фільтрація даних, неформальна освіта.